Biomedical Optics EXPRESS

Lightweight deep learning model incorporating an attention mechanism and feature fusion for automatic classification of gastric lesions in gastroscopic images

LINGXIAO WANG,[1,†] YINGYUN YANG,[2,†] AIMING YANG,[2,3,‡] AND TING LI[1,4,‡]

[1] *Institute of Biomedical Engineering, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin 300192, China*
[2] *Department of Gastroenterology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China*
[3] *yangaiming@medmail.com.cn*
[4] *liting@bme.cams.cn*
[†] Contributed equally
[‡] Co-corresponding authors

Abstract: Accurate diagnosis of various lesions in the formation stage of gastric cancer is an important problem for doctors. Automatic diagnosis tools based on deep learning can help doctors improve the accuracy of gastric lesion diagnosis. Most of the existing deep learning-based methods have been used to detect a limited number of lesions in the formation stage of gastric cancer, and the classification accuracy needs to be improved. To this end, this study proposed an attention mechanism feature fusion deep learning model with only 14 million (M) parameters. Based on that model, the automatic classification of a wide range of lesions covering the stage of gastric cancer formation was investigated, including non-neoplasm(including gastritis and intestinal metaplasia), low-grade intraepithelial neoplasia, and early gastric cancer (including high-grade intraepithelial neoplasia and early gastric cancer). 4455 magnification endoscopy with narrow-band imaging(ME-NBI) images from 1188 patients were collected to train and test the proposed method. The results of the test dataset showed that compared with the advanced gastric lesions classification method with the best performance (overall accuracy = 94.3%, parameters = 23.9 M), the proposed method achieved both higher overall accuracy and a relatively lightweight model (overall accuracy =95.6%, parameter = 14 M). The accuracy, sensitivity, and specificity of low-grade intraepithelial neoplasia were 94.5%, 93.0%, and 96.5%, respectively, achieving state-of-the-art classification performance. In conclusion, our method has demonstrated its potential in diagnosing various lesions at the stage of gastric cancer formation.

1. Introduction

Gastric cancer (GC) is the fourth leading cause of cancer-related deaths and the fifth most common cancer worldwide. It is reported that 770,000 deaths and 1.09 million (M) new cases of GC occurred worldwide in 2020 [1]. Some studies have shown that the prognosis of GC patients is closely related to the pathological stage, and the 5-year survival rate of patients with early gastric cancer(EGC) is over 90% [2], while the 5-year survival rate of patients with advanced gastric cancer(AGC) is only about 20% [3,4]. Notably, EGC develops from a series of precancerous lesions (gastritis, intestinal metaplasia (IM), low-grade intraepithelial neoplasia (LGIN), and high-grade intraepithelial neoplasia (HGIN)), and the more progressive the lesions, the higher the risk of patients developing EGC [5]. It is estimated that patients with IM and patients with LGIN are 10 and 25 times more likely to develop EGC than normal subjects,

#487456 https://doi.org/10.1364/BOE.487456
Journal © 2023 Received 10 Feb 2023; revised 11 Jun 2023; accepted 29 Jun 2023; published 11 Aug 2023

respectively [5–7]. Therefore, timely detection and accurate diagnosis of EGC and precancerous lesions can significantly reduce the mortality and incidence of EGC, which has become an urgent need for patients.

Magnification endoscopy with narrow-band imaging (ME-NBI) allows detailed observation of the morphological features of the gastric mucosal surface microstructure and has become the primary tool for the clinical examination of gastric mucosal lesions [8–11]. However, ME-NBI-based endoscopy still has some problems to solve. First, the features of gastric ME-NBI lesion images are complex, and magnification operation further affects the observation of features, making accurate diagnosis difficult even for experienced endoscopists [12,13]. Second, manual diagnosis of endoscopic images is labor-intensive, and the fatigue and lack of endoscopists may easily lead to diagnostic errors or missed diagnoses [14]. Therefore, the automated diagnosis of gastroscopic lesion images will bring tremendous clinical benefits, providing endoscopists with objective lesion assessment and helping them improve the accuracy and efficiency of diagnosing gastric lesions [15–19].

Machine learning techniques have advanced the state-of-the-art automatic detection of gastrointestinal endoscopic images in the past few years. Kanesaka et al [17]. proposed a method to classify gastric and non-gastric cancer based on gray-level covariance matrix (GLCM) features and eigenvector coefficient of variation extracted from ME-NBI images and obtained a accuracy of 96.3%. Van et al [20]. used a pre-trained support vector machine (SVM) and utilized high-definition endoscopic images' local texture and color features to classify early esophageal and non-early esophageal cancer with a system recall of 95.0%. Li et al [21]. used a uniform color local binary pattern algorithm to extract canonical color patterns from capsule endoscopy images and used the random forest (RF) to classify normal or diseased images with an accuracy of over 98.0%. However, these classification methods are highly dependent on the features defined by experts and cannot fit the diversity of features in reality, thus making it difficult to be extended to practical clinical.

Deep learning(DL) has recently been widely used in tasks such as automatic classification, segmentation, and localization of medical images [22–30]. Compared with traditional machine learning methods such as random forests and support vector machines, deep learning can automatically capture features in images with better flexibility and accuracy. Several scholars have demonstrated the applicability of deep learning in the automatic analysis of endoscopic images [31–38]. Horiuchi et al [39] used a GoogleNet-based transfer learning method to classify EGC and gastritis in gastric ME-NBI images with an accuracy of 85.3%. Liu et al [40] used a deep learning model to classify chronic gastritis (CG), LGIN, and EGC on gastric ME-NBI images with the recall of 92.0%, 92.0%, and 99.0%, respectively. Cho et al [41] developed an intelligent system based on three deep-learning models to automatically detect EGC and precancerous lesions in gastric WLI images with an average accuracy of 76.4%. Lui et al [42] used deep learning models to classify LGIN, HGIN, and cancer in gastric NBI images with an accuracy of 91.0%. Although these research groups realized the automatic detection of gastric lesions based on deep learning, there were no studies on intelligent recognition of various lesions covering the formation stage of gastric cancer (gastritis, IM, LGIN, HGIN, EGC) based on ME-NBI images. Moreover, the classification performance of precancerous lesions in existing studies needs to be improved. In addition, these methods were all based on transfer learning models, which may lead to overfitting, gradient disappearance, and even gradient explosion in the case of insufficient data.

The role of attention mechanisms in improving the performance of deep learning models has been demonstrated [43,44] and successfully applied in medical image analysis [45,46]. Feature fusion fuses multiple features with different properties into a new feature to obtain discrepancy information from the original feature, which can effectively improve the model's performance [47,48]. Inspired by F. Wang et al [44] and Dai et al [48], we proposed an efficient attention-mechanism feature fusion deep learning model. We used two trunk branches to obtain multiple semantic feature maps by using convolution units (based on separable convolution layers [49] or Inception units [50]) with convolution kernels of different sizes in different trunk branches. We fused these feature maps to get a new feature map, which was used as the input of the attention module. The attention branch of the attention module used an encoding-decoding structure to encode and decode the input feature fusion map multiple times and performed multi-scale feature fusion via skip connections to produce an image with stronger semantic information. We soft-weight the output of the attention branch to the output of the two trunk branches to guide the model for attention feature learning. To reduce the number of model parameters and computational complexity, we replaced the standard convolution kernel in the separable convolution layer with the dilated convolution kernel [51]. To solve the problem of class imbalance in the dataset, we introduced cost-sensitive learning into the model and weighted the loss function by cost penalty weight to improve the performance of feature learning. The main contributions of this paper are as follows:

(1) This paper proposed a deep-learning model of attention-mechanism feature fusion. As far as we know, this is the first time that the attention mechanism and feature fusion technology have been introduced into the automatic classification of gastric ME-NBI lesion images.

(2) A classification method based on the proposed model was developed to classify three major categories of gastric diseases (five subtypes of lesions covering the stage of gastric cancer formation). Under the condition that the number of model parameters (only 14 M) was close to half of the best-performing advanced gastric lesion classification model, our method achieved the highest overall accuracy, sensitivity, and specificity and the state-of-the-art classification performance for LGIN.

2. Methods

2.1. *Image acquisition and labeling*

The ME-NBI images for this study were retrospectively collected from patients who attended the Department of Gastroenterology at Peking Union Medical College Hospital from March 2014 to July 2021. The images were acquired using a GIF-H260Z endoscope (Olympus Medical Systems, Tokyo, Japan) and an EVIS LUCERA CV-290 endoscope system (Olympus Medical Systems) and saved at two resolutions of 1920 × 1080 pixels and 1440 × 1080 pixels as Joint Photographic Experts Group (JPEG) format image files. The study was approved by the Institutional Review Board of Peking Union Medical College Hospital, Beijing, China. And as a retrospective study, informed consent from patients was not required for this study. Images with poor quality due to defocus, mucus, under-inflation, and image blur were excluded from the study. Images lacking the pathological diagnosis were also excluded. The remaining images were included in the study. Black borders and text information containing patient privacy were removed from all images. The collected images covered five categories of gastritis, IM, LGIN, HGIN, and EGC. Since there is no universally accepted definition to distinguish intraepithelial neoplasia or cancer, and it is almost impossible to accurately differentiate between HGIN and EGC in actual clinical practice, HGIN can be classified into the EGC category [17, cho]. The five types of lesions are divided into three categories: EGC(including EGC and HGIN), LGIN, and non-neoplasm (including gastritis and IM). 4455 ME-NBI gastroscopy images from 1188 patients were finally collected, including 392 EGC images from 116 patients, 1351 LGIN images from 410 patients, and 2712 non-neoplasm images from 662 patients.

Two pathologists from Peking Union Medical College Hospital performed the pathological diagnosis, and patients with histologically confirmed EGC, LGIN, and non-neoplasm were included in this study. The pathological diagnosis of EGC was based on the tissue removed by

Biomedical Optics EXPRESS

endoscopic submucosal dissection or surgical excision. The pathological diagnosis of LGIN was made partly based on excised tissue and partly based on biopsy tissue. The pathological diagnosis of non-neoplasm was made based on biopsy tissue. Two endoscopists (with more than seven years of gastroscopy experience) from the Department of Gastroenterology of Peking Union Medical College Hospital classified the ME-NBI images. They first excluded the endoscopic images that did not match the pathological anatomical location in the pathology report to ensure that the captured location of the images included in the study was consistent with the endoscopists' suspected presence of an abnormal biopsy location or surgical location; after that, they classified the remaining images into three categories: EGC, LGIN, and non-neoplasm, in conjunction with the pathological diagnosis. The third endoscopist, who had more than ten years of experience in gastroscopy, examined the images on which the first two endoscopists disagreed and determined the final image category.

2.2. Dataset creation

The dataset was divided into training and test datasets. We selected 881 patients and obtained 3827 images for the training dataset, including 324 EGC images from 77 patients, 1151 LGIN images from 302 patients, and 2352 non-neoplasm images from 502 patients. We fit and optimized the model's parameters by 5-fold cross-validation on the training dataset. The dataset was divided into five groups with a patient-based random sampling method, and images from one patient were assigned to only one group. If a patient had different categories of lesions simultaneously, images of different categories of lesions from that patient might appear in other groups. Since the number of images of some lesions (e.g., EGC) in the training dataset was very small, to avoid overfitting the model, we increased the sample of the training dataset by image rotation transformation (±20°), flip transformation (vertical and horizontal), and other methods that did not affect the lesion characteristics of the images.

The remaining 628 images from 307 patients were used as the test dataset to evaluate the model's performance, including 68 EGC images from 39 patients, 200 LGIN images from 108 patients, and 360 non-neoplasm images from 160 patients. The overall median age of patients was 60 years, the range of 24-89, and the gender ratio between males and females was 189/118. To ensure the prospective nature of the test dataset, the endoscopy dates of patients in the test dataset (January 2021-July 2021) were not crossed with the examination dates of patients in the training dataset (February 2014-June 2020). Statistics for the training and test datasets are shown in Table 1.

Table 1. Demography of the gastric lesions dataset used in this study[a]

	Training dataset		Test dataset			
	No. of images	No. of patients	No. of images	No. of patients	Median age (range)	Sex (M/F)
EGC	324	77	68	39	69 (36-87)	26/13
LGIN	1151	302	200	108	59 (32-89)	70/38
Non-neoplasm	2352	502	360	160	51 (24-75)	93/67
Overall	3827	881	628	307	60 (24-89)	189/118

[a] EGC: early gastric cancer; LGIN: low-grade intraepithelial neoplasia; M: male; F: female

2.3. Attention-mechanism feature fusion model for lesions classification

Clinically, endoscopists usually diagnose gastric lesions based on the morphological features of different gastric mucosal lesions. Inspired by the diagnostic process of endoscopists, we proposed

Biomedical Optics EXPRESS

the attention-mechanism feature fusion convolution neural network(CNN) model (Fig. 1 (a)) for gastric ME-NBI image classification. We designed both the Xception-Attention-Feature-Fusion-Block(Xception-AFFB) and Inception-Attention-Feature-Fusion-Block (Inception-AFFB) attention feature fusion blocks as the main components of the model.

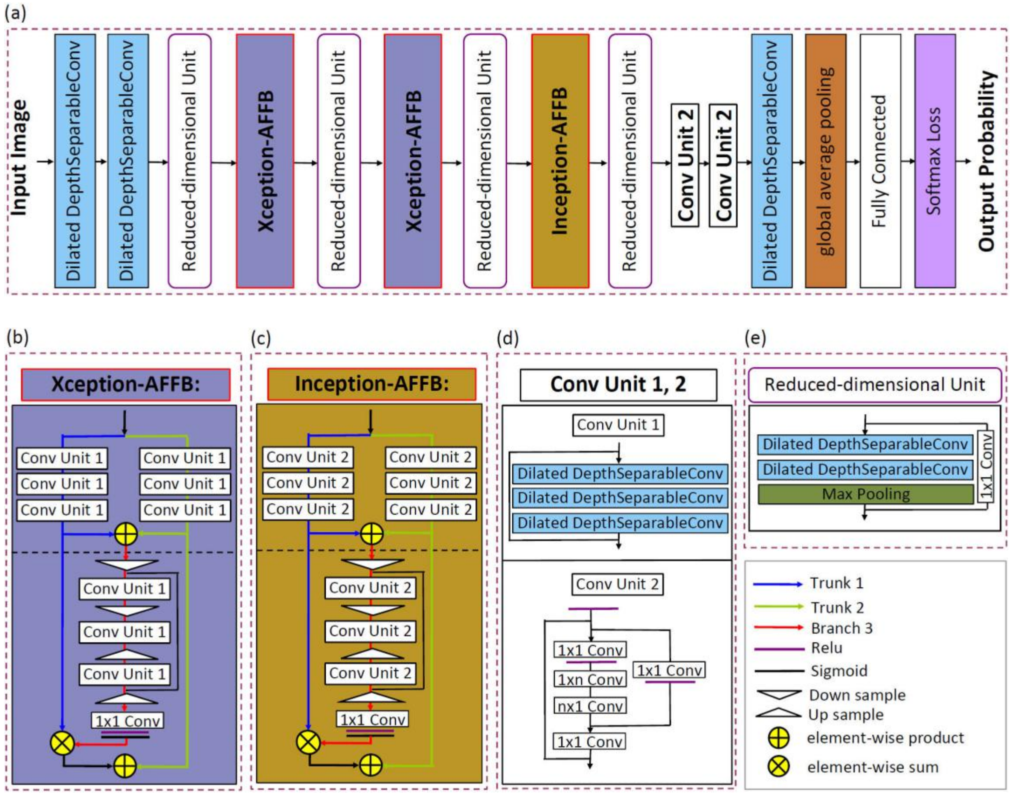


Fig. 1. Classification model architecture of gastric lesions and the proposed attention feature fusion blocks and convolution units. (a) Schematics of the proposed attention-mechanism feature fusion CNN model(Our). (b) The proposed Xception-Attention-Feature-Fusion-Block(Xception-AFFB). (c) The proposed Inception-Attention-Feature-Fusion-Block (Inception-AFFB). (d) Schematics of the convolution unit 1, 2(Conv Unit 1, 2).

Figure 1 (b) and (c) show the overview of the two blocks, both Xception-AFFB (Fig. 1 (b)) and Inception-AFFB (Fig. 1 (c)) consist of two parts: the multi-semantic feature fusion structure (Fig. 1 (b) and (c) above the horizontal black dashed line) and the attention module (Fig. 1 (b) and (c) below the horizontal black dashed line). Inspired by Dai et al. [48], our multi-semantic feature fusion structure consisted of two trunk branches, which were used to generate two feature maps with different semantic information. One branch (Trunk 1) generated low-semantic feature maps through convolution units with smaller convolution kernels, and the other branch (Trunk 2) generated high-semantic feature maps through convolution units with larger convolution kernels. We concatenated and fused the two semantic feature maps to generate a new feature map and fed it into the attention module. We used an encoding and decoding structure in the attention module's attention branch (Branch 3). The encoding structure consisted of multiple pooling layers and convolution units to generate low-resolution feature images with strong semantic information. The decoding structure consisted of multiple upsampling layers and convolution units to map low-resolution feature images to their original size and perform pixel-by-pixel classification. The skip connections were inserted in the encoding and decoding structure to connect lower-level detail features and higher-level semantic features to help the network obtain high-resolution feature images with high-level semantics. After the decoding structure, we used a 1 × 1 convolution layer (1 × 1 Conv) to achieve cross-channel information interaction, and we added a rectified linear unit(RELU) activation function after this convolution layer to enhance the learnable feature variation space. Finally, a sigmoid function was used to achieve a mixed attention constraint on channel and space. The feature maps output from Branch 3 was multiplied with the lower semantic feature maps output from Trunk 1 and then added with the higher semantic feature maps output from Trunk 2 to add soft weights to the trunk branches' feature maps of the network to achieve the intention of guiding the model feature learning.

In Xception-AFFB, Trunks1, 2, and Branch 3 each contained three convolution units 1 (Conv Unit 1, as shown in Fig. 1 (d)). The convolution layer in Conv Unit 1 was derived from the depth-separable convolution layer in Xception [49] network. The depth-separable convolution layer can effectively use the parameters without increasing the model's capacity and improve the model's performance. Therefore, we introduced the depth-separable convolution layer into Conv Unit 1; however, to reduce the number of model parameters, we replaced the standard convolution kernel in the depth-separable convolution layer with the dilated convolution kernel. The dilated convolution kernel allows the convolution kernel to increase the receptive field and feature representation capability without increasing the parameters by setting the dilated rate [49]. For example, the receptive field of a dilated convolution kernel of size 2 × 2 with a dilated rate of 2 is equivalent to the receptive field of a 3 × 3 standard convolution kernel, but the former takes up only four parameters, and the latter takes up nine parameters. We named the depth-separable convolution layer that replaced the standard convolution kernel as Dilated DepthSeparableConv layer and added the RELU activation function before each convolution layer and batch normalization after the convolution layer. We stacked three such convolution layers to form Conv Unit 1. We used the dilated convolution kernel of size 2 × 2 with a dilated rate of 2 (equivalent to the receptive field of a standard convolution kernel of size 3 × 3) in Conv Unit 1 of Trunk 1 and the dilated convolution kernel of size 3 × 3 with a dilated rate of 2 (equivalent to the receptive field of a standard convolution kernel of size 5 × 5) in Conv Unit 1 of Trunk 2. Compared with Conv Unit 1 in Trunk 1, the convolution kernel in Conv Unit 1 in Trunk 2 has a larger receptive field and can acquire feature maps with more global features and higher semantic levels. Since the output of Branch 3 was multiplied and weighted with the output of Trunk 1, we used the same size convolution kernel in Conv Unit 1 of Branch 3 as in Conv Unit 1 of Trunk 1.

In Inception-AFFB, Trunks 1, 2, and Branch 3 each contained three convolution units 2 (Conv Unit 2, as shown in Fig. 1 (d)), which used the Inception block of Inception-ResNet V2 [50]. Inception block achieves multi-level feature fusion through multi-branch parallel structure; at the same time, it factorizes a convolution layer with n × n convolution kernel size into two convolution layers with 1×n (1×n Conv) and n × 1 (n × 1 Conv) convolution kernel size, reduces the number of parameters from (n × n) to (1×n + n × 1), and extends the model depth. The application of these techniques allows Inception blocks to achieve better performance with lower computational costs. Like Trunks1 and 2 in Conv Unit 1, we used convolution kernels with n = 3 and n = 5 in Conv Unit 2's Trunks1 and 2, respectively, to generate multi-semantic feature maps; and used the same size convolution kernels in Branch 3's Conv Unit 2 as in Trunk 1's Conv Unit 2.

The proposed attention-mechanism feature fusion convolution neural network model (Our) architecture is shown in Fig. 1 (a). We stacked multiple attention feature fusion blocks to gradually refine the attention to the trunk feature map with incremental features. Specifically, the model contained three Dilated DepthSeparableConv layers at the head and tail feature extraction positions; one Xception-AFFB at the low and middle feature extraction positions, respectively; one Inception-AFFB at the high feature extraction position; and two Conv Unit 2 at the top feature extraction position which the convolution kernels with n = 3 because of the small feature map size at this position. In addition, we used the reduced-dimensional unit(as shown in Fig. 1 (e)) as the transition structure between the attention feature fusion blocks, which consisted of two Dilated DepthSeparableConv layers (with 2 × 2 convolution kernel size and dilated rate of 2) and one Maxpooling layer. Our reduced-dimension unit can effectively balance the feature map's maximum pooling operation and feature representation bottleneck and save the occupied parameters. We used a global average pooling layer, a fully connected layer, and a Softmax loss function to form the classifier of the model, which outputs the probability values of the input images belonging to the three gastric lesions. In addition, We replaced all the attention feature fusion blocks in Fig. 1 (a) with Xception-AFFB to obtain Xception-AFFCNN and replaced all the attention feature fusion blocks in Fig. 1 (a) with Inception-AFFB to obtain Inception-AFFCNN. We evaluated the classification performance of the proposed models(Xception-AFFCNN, Inception-AFFCNN, and Our) on gastric lesions datasets and compared them with the benchmark models.

2.4. Class imbalance handling method

Deep learning-based medical image analysis tasks usually suffer from the dataset class imbalance problem, i.e., one class contains significantly more samples than the other classes, and our dataset suffers from this problem. For example, we have 2712 non-neoplasm images, but the number of images for LGIN and EGC is only 1351 and 392, respectively. Class imbalance can make the model feature learning highly biased towards the majority class and affect the accuracy of classification results. Cost-sensitive learning [52] is a common approach to solving this problem, which usually minimizes the cost and loss function value of misclassification by using a smaller penalty term for the majority class and a larger penalty term for the minority class, thus rebalance the classes. Therefore, we introduced cost-sensitive learning into the model's loss function to address the dataset class imbalance problem. We designed penalty weight values for each category, as shown in Eq. (1):

$$W_j = \frac{N_{total}}{C \cdot N_j} \tag{1}$$

In the above equation, j represents the category ($j = 1,2,3$), N_{total} represents the total number of samples, C represents the number of categories ($C = 3$), N_j represents the number of samples in category j, W_j represents the penalty weight value of this category. Then we introduced the weight value into the loss function to obtain the weighted loss function, which was calculated as shown in Eq. (2):

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} W_j \cdot Y_{ij} \cdot \log(P_{ij}) \tag{2}$$

In the above equation, N represents the number of batch samples, i represents the batch sample order, Y_{ij} represents the true sample label, and P_{ij} represents the predicted probability value. By the above method, the model will consider the misclassification cost and adjust the loss values of different classes to rebalance the classes during the training process.

2.5. Model training

We trained each model by 5-fold cross-validation on the training dataset. The training dataset was divided into five groups at the patient level, and each model was cross-validated five times; in each cross-validation, one different group was used for model validation, and the remaining four groups were used for model training. The learning rate was 1E-4, batch size was 8, and Adam was used as the optimizer. Since the model's performance in the validation group was not further improved, the number of training epochs was 150. All models used the same training parameters. The average results of 5-fold cross-validation for each model were calculated on independent test dataset. All models were trained and tested on an AMD Ryzen 7-1700X eight-core processor CPU(central processing units) and a GeForce GTX 1080 Ti GPU(graphics processing units).

2.6. Statistical analysis

The performance of each model was evaluated with per-category accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with 95% confidence intervals (CI). In addition, the receiver operating characteristic (ROC) curve was also used to evaluate the comprehensive classification ability of the model. The area under the curve(AUC) was automatically calculated based on the ROC curve, and the range was distributed between 0 and 1, higher values indicated better comprehensive classification performance of the model. The calculation of each evaluation metric was shown in Eq. (3)-(7):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{4}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{5}$$

$$PPV = \frac{TP}{TP + FP}, \tag{6}$$

$$NPV = \frac{TN}{TN + FN}, \tag{7}$$

Where TP indicates True Positive, FP indicates False Positive, TN indicates True Negative, and FN indicates False Negative. The overall accuracy (O-ACC), sensitivity (O-SE), specificity (O-SP), PPV (O-PPV), NPV (O-NPV), and AUC (O-AUC) were obtained by summing and averaging the same metric for all categories.

3. Results

3.1. Model evaluation on the gastric lesions dataset

In this section, we evaluated the performance of our model on the gastric test dataset. We performed ablation studies and comparison experiments with other relevant state-of-the-art methods, respectively. The overall accuracy (O-ACC), sensitivity (O-SE), specificity (O-SP), PPV (O-PPV), NPV (O-NPV), and AUC (O-AUC) were used to evaluate the performances of each method.

3.1.1. Ablation studies

We first conducted ablation studies to show the proposed model's effectiveness. We compared the classification performance of the benchmark models(Xception, Inception-ResNet V2) and the proposed models(Xception-AFFCNN, Inception-AFFCNN, and Our) on the gastric test dataset. To make a fair comparison, each model applies hyperparameters to achieve optimal performance; the hyperparameters of Xception and Inception-ResNet V2 are the same (learning rate = 1E-4, batch size = 8, and epochs = 100), and the proposed model uses the same hyperparameters (learning rate = 1E-4, batch size = 8, and epochs = 150). The classification results are shown in Table 2. In the multi-classification comparison of non-neoplasm, LGIN, and EGC, our model(Our) obtained the optimal values (the bold fonts in Table 2) for O-ACC, O-SE, O-SP, O-PPV, O-NPV, and O-AUC. The proposed Xception-AFFCNN achieved sub-optimal classification performance. Compared with the benchmark model Xception, the Xception-AFFCNN achieved performance gains on O-ACC, O-SE, O-SP, O-PPV, O-NPV, and O-AUC by 0.5%, 0.9%, 0.8%, 0.5%, 0.4%, and 0.3%, respectively. The results showed that the proposed models(Xception-AFFCNN, Our) based on attention feature fusion architecture effectively improve gastric lesions' classification performance. Compared with the benchmark model Inception-ResNet V2, the proposed Inception-AFFCNN did not significantly improve classification performance. We also count the number of parameters of each model to indicate the complexity of the model. The number of parameters for Our is 14 M, far less than Xception's parameters and nearly a quarter of Inception-ResNet V2's parameters.

Table 2. Comparison of ablation studies on gastric lesions dataset[a]

Methods	O-ACC	O-SE	O-SP	O-PPV	O-NPV	O-AUC	P(M)
Inception-ResNet V2 [50]	93.7 (93.1, 94.3)	85.1 (82.9,87.3)	93.7 (92.9,94.5)	91.9 (90.9, 92.9)	95.1 (94.8,95.4)	98.0 (97.7, 98.3)	55.9
Xception [49]	94.6 (93.9, 95.3)	89.5 (88.5,90.5)	94.6 (93.9,95.3)	91.4 (89.7, 93.1)	95.7 (94.9,96.5)	98.2 (97.9, 98.5)	22
Inception-AFFCNN	92.7 (92.0, 93.4)	88.4 (83.4,93.4)	93.2 (91.5,94.9)	88.7 (86.2, 89.6)	94.6 (92.4,96.8)	98.1 (97.9, 98.3)	13.3
Xception-AFFCNN	95.1 (94.3, 95.9)	90.4 (88.9,91.9)	95.4 (94.9,96.6)	91.9 (87.2, 98.8)	96.1 (94.7,97.5)	98.5 (98.2, 98.8)	16.2
Our	**95.6** **(94.6, 96.6)**	**92.8** **(89.7,95.9)**	**96.2** **(95.3,97.1)**	**93.6** **(92.1, 95.1)**	**96.6** **(95.3,97.9)**	**98.9** **(98.7, 99.1)**	14.0

[a] O-ACC:overall accuracy; O-SE:overall sensitivity; O-SP:overall specificity; O-PPV:overall positive
[a] predictive value; O-NPV:overall negative predictive value; O-AUC:overall area under the curve;
[a] P: parameters; M: million.

The depth-separable convolution layer in Xception [49] is an extreme case of the convolution layer in the Inception module [50]. The former has better performance gain [49]. This phenomenon also exists in our experimental results (classification results of Xception and Inception-ResNet V2 in Table 2). This is due to the efficient use of parameters in the depth-separable convolution layer. In fact, the depth-separable convolution layer does not reduce the number of parameters for Xception. On the contrary, at the same depth position of the network, the number of parameters occupied by the depth-separable convolution layer is much higher than that occupied by the convolution layer of the Inception module. Taking our designed Inception-AFFB and Xception-AFFB as examples, the two blocks have the same structural framework and number of convolution units. The difference is that Conv unit 2 of Inception-AFFB contains four factorization convolution layers. However, Conv unit 1 of Xception-AFFB contains three depth-separable convolution layers, less than the number of convolution layers in Conv unit 2 of Inception-AFFB. However, when the Inception-AFFB at the high-level feature extraction in Fig. 1 (a) is replaced with Xception-AFFB (get our proposed Xception-AFFCNN), the model's parameters will increase by 2.2 M. It indicates that the number of parameters occupied by Xception-AFFB at the same location of the model is much higher than that occupied by Inception-AFFB. The results in Table 2 showed that the classification result of Inception-AFFCNN was not the best, but it still maintained a good performance when the number of parameters was reduced to 13.3 M. Therefore, based on the above analysis, we designed our model(Our), as shown in Fig. 1 (a), to better balance the number of parameters and classification performance. We used Xception-AFFB in the low-level and medium-level feature extraction positions to improve the model's performance and used Inception-AFFB with less parameter occupation to control the model parameters in the high-level feature extraction positions. As seen from the results in Table 2, our model (Our) had better classification performance than Xception-AFFCNN using only Xception-AFFB and Inception-AFFCNN using only Inception-AFFB. The number of model parameters was also controlled (14.0 M vs. 16.2 M vs. 13.3 M). The results showed that compared with the benchmark models (Xception, Inception-ResNet V2) and the model designed by us (Xception-AFFCNN, Inception-AFFCNN), our model(Our) achieved higher classification performance and achieved lightweight of the model. In the following research, we only analyze our model(Our), and our method in the following paper refers to the classification method based on our model(Our).

3.1.2. Comparisons with the state-of-the-art methods

Then, we compared the classification performance of our method with that of three other advanced methods for classifying gastric lesions [34–35,40], including the classification method proposed by Zhang et al [34] for chronic atrophic gastritis; the classification method proposed by Li et al [35]. for EGC based on ME-NBI images; and the classification method designed by Liu et al [40] for precancerous gastric lesions based on ME-NBI images. To ensure the best performance of each method, the input image size of all methods was kept the same as the original network, with the input image size of 299 × 299 for Li et al.'s method, 224 × 224 for Zhang et al.'s, Liu et al.'s, and our method, and other training conditions were kept the same.

Table 3 summarizes the overall classification results for the three gastric lesions (non-neoplasm, LGIN, and EGC) of the four methods in the gastric test dataset. As can be seen from Table 3, our method achieved optimal values (values in bold) for O-ACC, O-SE, O-SP, O-PPV, O-NPV, and O-AUC, and were 1.3%, 4.2%, 1.9%, 2.0%, 1.2%, and 0.8% higher than the sub-optimal values, respectively. This indicated that our method outperformed the other three methods in terms of overall classification performance. In terms of the number of parameters, the number of parameters of our model was higher than that of Zhang et al.(14 M vs. 8.1 M) and much less than that of Li et al.(23.9 M) and Liu et al.(26.0 M). Regarding the time complexity, the average test time of our method was 0.73 ms per image, while the average test times of Zhang et al., Li et al., and Liu et al. were 0.51 ms, 0.78 ms, and 0.89 ms per image, respectively. Although the number of parameters and time complexity of our method was higher than those of Zhang et al.'s method, the classification performance was significantly better than that of Zhang et al., with O-ACC, O-SE, O-SP, O-PPV, O-NPV, and O-AUC improving by 1.9%, 7.7%, 2.5%, 1.7%, 1.5%, and 0.9%, respectively. The comparison results showed that our method achieved a better balance among classification performance, the number of parameters, and time complexity than the other three advanced classification methods for gastric lesions.

Table 3. Comparison with other state-of-the-art methods on gastric lesions dataset[a]

Methods	O-ACC	O-SE	O-SP	O-PPV	O-NPV	O-AUC	P(M)
Zhang Y, et al. [34]	93.0	85.3	93.1	89.8	94.5	96.8	8.1
	(92.2, 93.8)	(83.4,87.2)	(92.2,94.0)	(87.7, 91.9)	(93.7, 95.3)	(96.5, 97.1)	
Li, et al. [35]	94.3	88.6	94.3	91.6	95.4	98.1	23.9
	(93.6, 95.0)	(87.9,89.3)	(93.6,95.0)	(89.7, 93.5)	(94.6, 96.2)	(97.9, 98.3)	
Liu, et al. [40]	94.0	86.1	94.2	90.9	95.3	97.7	26.0
	(93.0, 95.0)	(83.3,88.9)	(93.1,95.3)	(89.3, 92.5)	(94.5, 96.1)	(97.4, 98.0)	
Our	**95.6**	**92.8**	**96.2**	**93.6**	**96.6**	**98.9**	14.0
	(94.6, 96.6)	**(89.7,95.9)**	**(95.3,97.1)**	**(92.1, 95.1)**	**(95.3,97.9)**	**(98.7, 99.1)**	

[a] O-ACC:overall accuracy; O-SE:overall sensitivity; O-SP:overall specificity; O-PPV:overall positive
[a] predictive value; O-NPV:overall negative predictive value; O-AUC:overall area under the curve;
[a] P: parameters; M: million.

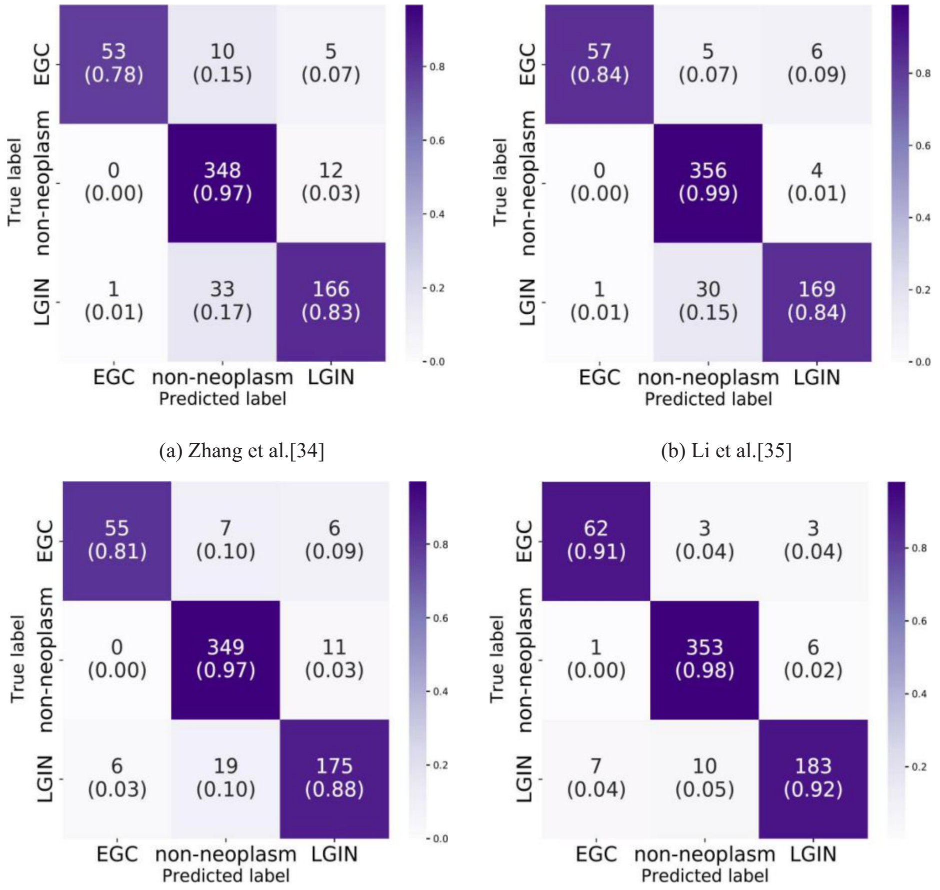
In addition, we also counted the per-category accuracy, sensitivity, specificity, PPV, NPV, and AUC for each method on the test dataset(Table S1 in Supplement 1). The results showed that in the case of LGIN, our method achieved optimal values for accuracy, sensitivity, specificity, PPV, NPV, and AUC. In the case of non-neoplasm detection, our method also achieved better values for accuracy, specificity, PPV, NPV, and AUC than the corresponding metrics of other methods, except for sensitivity. In the case of EGC, our method achieved optimal values for accuracy, sensitivity, NPV, and AUC, while specificity and PPV were lower than the values of other methods.

Figure 2 shows the confusion matrixes for each method on the gastric test dataset, which provide a more visual representation of the correlation between the predicted and actual labels and the distribution of the number of predicted labels for each method. We used the per-category true positive(TP) rate to describe the confusion matrix results of each method. As can be seen from Fig. 2, in the case of EGC, Zhang et al.'s method obtained the lowest TP rate of 0.78 (as shown in Fig. 2 (a)); Li et al.'s method and Liu et al.'s method obtained the TP rate of 0.84 (as shown in Fig. 2 (b)) and 0.81 (as shown in Fig. 2 (c)), respectively; our method achieved the highest TP rate of 0.91(as shown in Fig. 2 (d)). 15% of EGC were misclassified as non-neoplasm in Zhang et al.'s method, while only 4% of EGC were misclassified as non-neoplasm in our method. In the case of non-neoplasm, Li et al.'s method achieved the highest TP rate of 0.99, while Zhang et al.'s method had the lowest TP rate of 0.97; in these two methods, 1% and 3% of non-neoplasm were misclassified as LGIN, respectively; while our method had a TP rate of 0.98 and 2% of non-neoplasm were misclassified as LGIN. In the case of LGIN, our method achieved the highest TP rate of 0.92; Zhang et al.'s method had the lowest TP rate of 0.83. The results in Fig. 2 showed that Our method had the highest TP rate for EGC and LGIN and was only 1% lower than the highest TP rate for non-neoplasm (0.98 vs. 0.99). The comprehensive results showed that our method was more suitable for detecting gastric lesions in ME-NBI images.

In addition, to further evaluate the comprehensive classification ability of each method, the ROC curves of each method on EGC, LGIN and non-neoplasm were drawn, and AUC values were obtained, as shown in Fig. 3. The AUC of our method for EGC, LGIN and non-neoplasm were 0.998, 0.986 and 0.990, respectively, which were superior to those of other methods. The results showed that our method showed better comprehensive classification performance for each type of gastric lesion.

3.2. Model generalization ability on the UCSD dataset

To validate the generalization ability of our method on other datasets, we evaluated the classification performance of our method on the publicly available retinal spectral-domain optical coherence tomography (SD-OCT) dataset (UCSD, University of California San Diego) and compared it with three other methods (Zhang et al [34], Li et al [35], and Liu et al [40]). Optical detection imaging plays an important role in the clinic [53]. Like endoscopic imaging, SD-OCT uses light to capture micron-resolution optical cross-sections of human tissues and assembles them into three-dimensional volumetric images [54]. Due to its high resolution and real-time imaging, have been widely used in various medical applications [55–56], especially in ophthalmic diagnosis. The UCSD dataset provides 109,309 retinal SD-OCT 2D images in Tag Image File Format (TIFF) format, containing diabetic macular edema (DME), choroidal neovascularization (CNV), DRUSEN, and NORMAL. The image labels are strictly labeled by ophthalmologists and used as the gold standard for the classification task. All images were divided into the training dataset and the test dataset. The training dataset had 108,309 images, including 37,205 CNV images, 11,348 DME images, 8,616 DRUSEN images, and 51,140 NORMAL images; the test dataset had 1,000 images, with 250 images each for CNV, DME, DRUSEN, and NORMAL. We trained the models of the four methods(Zhang et al [34], Li et al [35], Liu et al [40], and Our) on the training dataset using a 10-fold cross-validation method. We divided the training dataset into ten groups and performed ten times of cross-validation. In each cross-validation, a different group was used as the validation group to supervise the optimization of model parameters, and the remaining



(a) Zhang et al.[34]

(b) Li et al.[35]

(c) Liu et al.[40]

(d) Our

Fig. 2. Confusion matrixes of the gastric lesions classification for four methods on test dataset. (a) Results obtained by the method of Zhang et al [34]. (b) Results obtained by the method of Li et al [35]. (c) Results obtained by the method of Liu et al [40]. (d) Results obtained by our method. EGC: early gastric cancer; LGIN: low-grade intraepithelial neoplasia.

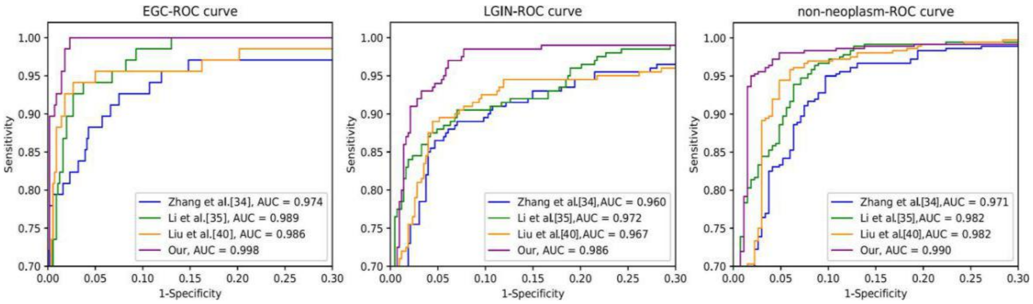


Fig. 3. Per-category ROC curves and AUC comparison of the four methods on the test dataset. ROC: receiver operating characteristic; AUC: area under the curve. EGC: early gastric cancer; LGIN: low-grade intraepithelial neoplasia.

nine groups were used as the training group to fit the model parameters. Adam was used as the optimizer in training with an initial learning rate of 1E-4 and a batch size of 8. Since the performance of the validation group was no longer improved, the number of training epochs was 50, and all models were trained with the same training parameters. All models' average results of 10-fold cross-validation were counted and compared on the independent test dataset.

Table 4 summarized the O-ACC, O-SE, O-SP, O-PPV, and O-NPV for each method on the UCSD independent test dataset, and the bold font represented the optimal values. Our method achieved the optimal O-ACC, O-SE, O-SP, O-PPV, and O-NPV values, and 0.5%, 1.0%, 0.3%, 0.7%, and 0.3% higher than the sub-optimal values, respectively. This indicated that the overall classification performance of our method on retinal OCT images exceeded that of other comparative methods. In addition, we also summarized the per-category classification performance of each method for CNV, DME, DRUSEN, and NORMAL(Table S2 in Supplement 1). The accuracy of our method for CNV, DME, DRUSEN, and NORMAL all reached the optimal value (as shown in bold font). Most of the other evaluation metrics also achieved the optimal value. In terms of time complexity, the per-image prediction time of our method on the UCSD independent test dataset was 0.66 ms, while the per-image prediction times of Zhang et al.'s, Li et al.'s, and Liu et al.'s methods were 0.67 ms, 1.2 ms, and 0.88 ms, respectively. This indicated that the time complexity of our method was lower than that of the other methods. The results showed that our method outperformed other comparison methods regarding classification performance and time complexity on the retinal OCT image dataset.

Table 4. Comparison with other state-of-the-art methods on the UCSD test dataset[a]

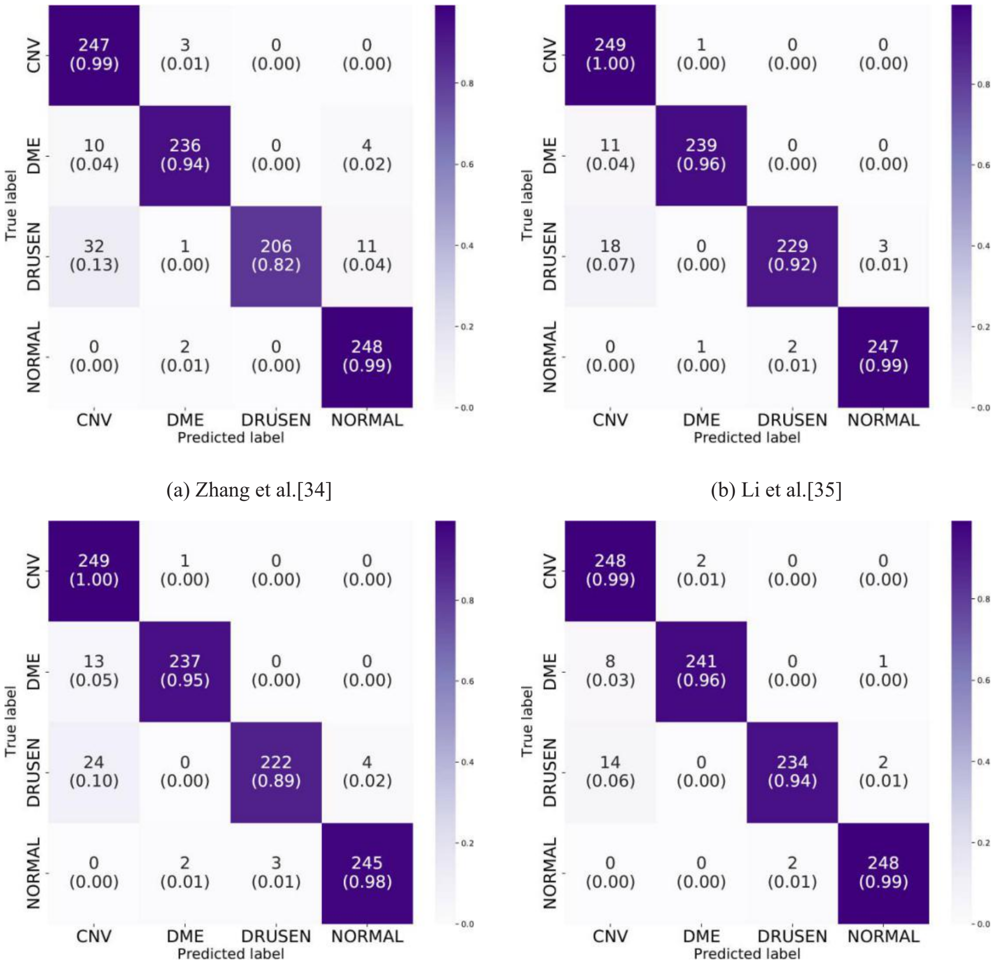
Methods	O-ACC	O-SE	O-SP	O-PPV	O-NPV
Zhang Y, et al. [34]	97.1	94.1	98.0	94.7	98.1
	(96.4, 97.8)	(92.4, 95.8)	(97.4, 98.6)	(92.8, 96.6)	(97.5, 98.7)
Li, et al. [35]	98.0	96.0	98.7	96.4	98.7
	(97.3, 98.7)	(94.4, 97.6)	(98.1, 99.3)	(95.0, 97.8)	(98.3, 99.1)
Liu, et al. [40]	97.9	95.7	98.5	96.0	98.6
	(97.2, 98.6)	(94.1, 97.3)	(97.9, 99.1)	(94.6, 97.4)	(98.2, 99.0)
Our	**98.5**	**97.0**	**99.0**	**97.1**	**99.0**
	(98.1, 98.9)	**(96.4, 97.6)**	**(98.9, 99.1)**	**(96.5, 97.7)**	**(98.9, 99.1)**

[a] O-ACC:overall accuracy; O-SE:overall sensitivity; O-SP:overall specificity; O-PPV:overall
[a] Positiv predictive value.

In addition, we used the confusion matrixes to visualize the distribution of predicted labels for each method on the UCSD test dataset, as shown in Fig. 4. As can be seen from Fig. 4, in the CNV case, Zhang et al.'s method achieved a TP rate of 0.99 (as shown in Fig. 4 (a)); Li et al.'s and Liu et al.'s methods achieved the highest TP rate, both reaching 1 (as shown in Fig. 4 (b), Fig. 4 (c), respectively); our method achieved a TP rate of 0.99 (as shown in Fig. 4 (d)), which lower than the maximum TP rate. However, our method gained the highest TP rates for the DME, DRUSEN, and NORMAL, respectively, and were 2%, 12%, and 1% higher than the lowest values. The comprehensive results showed that our method had good generalization ability on the retinal OCT image dataset.

3.3. Visualizing model decisions

To understand the diagnostic process of the model, we visualized the model's decision process using class activation mapping (CAM) to see whether the model's diagnosis was based on key clinical features. One case of EGC, LGIN, and non-neoplasm ME-NBI images with model



(a) Zhang et al.[34]

(b) Li et al.[35]

(c) Liu et al.[40]

(d) Our

Fig. 4. Confusion matrixes of the retinal lesions images classification for four methods on the UCSD test dataset. (a) Results obtained by the method of Zhang et al [34]. (b) Results obtained by the method of Li et al [35]. (c) Results obtained by the method of Liu et al [40]. (d) Results obtained by our method. CNV: choroidal neovascularization; DME: diabetic macular edema.

predictions consistent with the actual category was randomly selected from the gastric lesion test dataset, respectively. Using the class activation heat map generated by the CAM visualized the main regions of interest of the model to the image. The original ME-NBI images and the class activation heat map are shown in Fig. 5.

In Fig. 5, the first row from left to right showed the original EGC, LGIN, and non-neoplasm images, respectively. The yellow rectangular region in the image was the identification region of lesions marked by experts; hallmarks of these diseased areas were visible (EGC with obvious boundary line and irregular curved blood vessels, LGIN with epithelial atypia and mucosal structural changes, non-neoplasm with intestinal epithelial cells). The second row showed the class activation heat map corresponding to the original ME-NBI image in the first row. The orange and red regions in the class activation heat map indicated the regions with the most

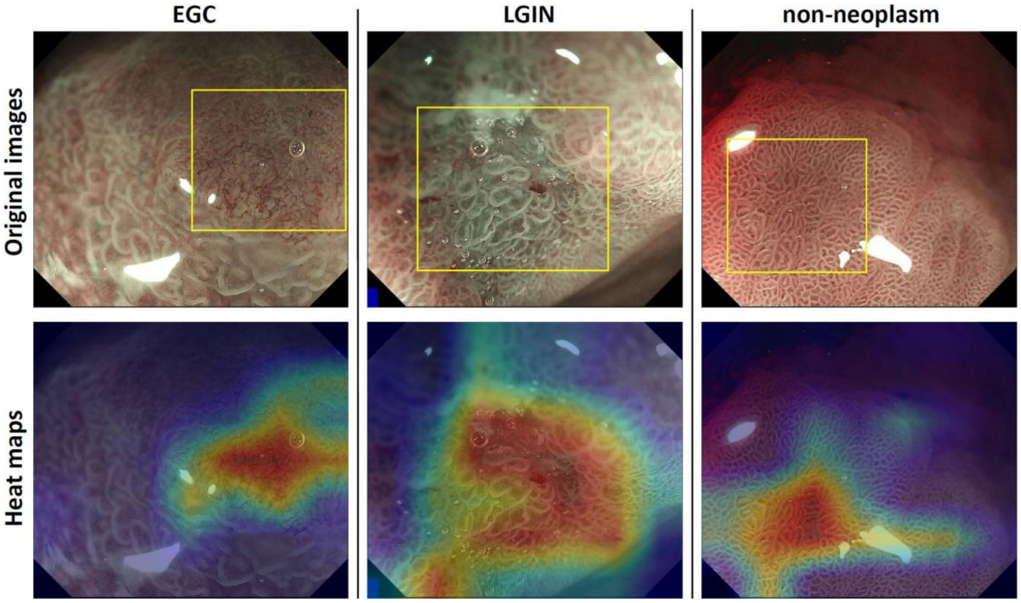


Fig. 5. Visualize the model classification decision process using the class activation heat map. The first line is the original image, and the yellow rectangular box areas are the lesion regions annotated by the experts. The second line is the corresponding heat map, where the orange-red area indicates a significant influence on the model decision, while the blue and green regions indicate a limited effect on the model decision. EGC: early gastric cancer; LGIN: low-grade intraepithelial neoplasia.

significant influence on the model's decision, and the features of these regions were the main basis for the model to make a classification decision. In contrast, the blue and green regions in the class activation heat map indicated limited influence on the model's decision. The class activation heat map showed that our model could accurately identify the clinical features of ME-NBI images of various gastric mucosal lesions and make a classification diagnosis, and the diagnostic basis was consistent with that of the experts. The visualization results showed that our model has good reliability.

4. Discussion

In this study, we proposed an attention-mechanism feature fusion deep learning model, based on which we established an automatic classification method that can detect a wide range of lesions covering the formation stage of gastric cancer from ME-NBI images. These lesions were classified as non-neoplasm(including gastritis and IM), precancerous neoplasm (LGIN), and EGC(including HGIN and EGC). As far as we know, this is the first time that the attention mechanism and feature fusion technology have been introduced into the automatic detection of gastric lesions.

The results showed that our method achieved the highest overall accuracy, sensitivity, specificity, PPV, NPV, and AUC, outperformed the classification performance of the benchmark models and other advanced methods, and outperformed the existing studies [41]. The detection accuracy and sensitivity of LGIN were 94.5% and 93.0%, respectively, achieving the most advanced classification performance.

Currently, the automatic classification methods of gastric lesions based on deep learning mainly adopt transfer learning methods. Compared with these methods [34–35,40], our approach achieved the optimal overall classification performance on the gastric lesions test dataset. We designed and built a lightweight model with only 14.0 M parameters. We used the depth-separable convolution layer [49] as the basic convolution layer for the low and medium levels of the network and replaced the standard convolution kernel with the dilated convolution kernel to reduce the model parameters. Furthermore, we used the factorization convolution layer [50], a convolution layer with relatively low parametric occupancy, as the basic convolution layer for the high levels of the network. Compared with the methods of Li et al [35] and Liu et al [40] (with 23.9 M and 26.0 M parameters, respectively), our method's number of model parameters was significantly reduced, while the overall accuracy was improved by 1.3% and 1.6%, respectively. Our method's prediction time per image was 0.73 ms, which was lower than the time complexity of Li et al [35] and Liu et al [40] (0.78 ms and 0.89 ms respectively). Our method was characterized by high precision, lightweight, and real-time, demonstrating the extensive potential of intelligent clinical diagnosis.

Compared with our previous method [57], instead of only obtaining a single feature map from a single trunk branch and feeding it to the attention module, the new method used multiple trunk branches to generate multiple semantic feature maps and the feature fusion map as input to the attention module. In addition, we designed a new attention module. We used multiple encoding and decoding layers in the attention branch and added the convolution units and skip-connection to carry out more sufficient convolution operations on the input feature fusion map. Furthermore, we weighted and fused the feature maps output by the attention branch to the feature maps with different semantic information output by multiple trunk branches instead of only weighing the output of a single trunk branch. Compared with our previous study [57], the proposed method significantly improved the overall accuracy and the per-category accuracy of gastric lesions.

Despite these advantages, there are some limitations to this study. Firstly, the data used in our study were from the same center, and the equipment, imaging settings, and image formats of different centers may have some influence on the classification performance; therefore, in the subsequent study, we will collect data from different centers and analyze them. Secondly, we used high-quality images in this study instead of low-quality ones (such as blurred areas and weak light). We will add more low-quality images in subsequent studies to enhance the robustness and universality of our method for low-quality images. In addition, this study only included gastric lesions; after collecting more patients and endoscopic images, we will include lesions such as esophagitis and early esophageal cancer into our lesion classification system to increase their clinical application.

5. Conclusion

This study proposes a gastric lesion classification method based on an attention-mechanism feature fusion deep learning model. Compared with previous gastric lesion classification models, our proposed model uses feature fusion techniques to enhance the recognition of ME-NBI image features and combines attention mechanisms to focus the model on lesion regions with significant features in the images. Our model achieves efficient classification with only 14 million parameters and a good balance of accuracy, time complexity, and the number of parameters. The proposed method is validated to outperform other state-of-the-art classification methods on both the gastric lesion and the retinal OCT datasets.

Funding. National Natural Science Foundation of China (81971660); Chinese Academy of Medical Science health innovation project (2021- I2M-1-058, 2021-I2M-1-042, 2022-I2M-C&T-A-005); Sichuan Science and Technology Program (2021YFH0004); Tianjin Outstanding Youth Fund Project (20JCJQIC00230); CAMS Innovation Fund for Medical Sciences (CIFMS) (2022-I2M-C&T-B-012); National High-Level Hospital Clinical Research Funding (2022-PUMCH-B-024).

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin. **71**(3), 209–249 (2021).
2. H. Suzuki, I. Oda, S. Abe, M. Sekiguchi, G. Mori, S. Nonaka, and Y. Saito, "High rate of 5-year survival among patients with early gastric cancer undergoing curative endoscopic submucosal dissection," Gastric Cancer **19**(1), 198–205 (2016).
3. H. Katai, T. Ishikawa, K. Akazawa, Y. Isobe, I. Miyashiro, and I. Oda, Registration Committee of the Japanese Gastric Cancer Association, "Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese Gastric Cancer Association (2001–2007)," Gastric Cancer **21**(1), 144–154 (2018).
4. H. J. Chun, B. Keum, J. H. Kim, and S. Y. Seol, "Current status of endoscopic submucosal dissection for the management of early gastric cancer: a Korean perspective," World J. Gastroenterol. **17**(21), 2592 (2011).
5. P. Correa and M. B. Piazuelo, "The gastric precancerous cascade," J. Dig. Dis. **13**(1), 2–9 (2012).
6. D. Li, M. C. Bautista, S. F. Jiang, P. Daryani, M. Brackett, M. A. Armstrong, and U. Ladabaum, "Risks and predictors of gastric adenocarcinoma in patients with gastric intestinal metaplasia and dysplasia: a population-based study," Official J. Am. College Gastroenterol. **111**(8), 1104–1113 (2016).
7. W. K. Leung and J. J. Y. Sung, "Intestinal metaplasia and gastric carcinogenesis," Alimentary Pharmacol. Therapeut. **16**(7), 1209–1216 (2002).
8. K. Yao, "Clinical application of magnifying endoscopy with narrow-band imaging in the stomach," Clin. Endosc. **48**(6), 481–490 (2015).
9. M. Kaise, M. Kato, M. Urashima, Y. Arai, H. Kaneyama, Y. Kanzazawa, and H. Tajiri, "Magnifying endoscopy combined with narrow-band imaging for differential diagnosis of superficial depressed gastric lesions," Endosc. **41**(04), 310–315 (2009).
10. O. M. Canales, J. Miyagui, J. Takano, and G Aliaga, "Tu2019 comparison between light blue crest and marginal turbid band for diagnosis of gastric intestinal metaplasia using narrow band imaging and magnifying endoscopy," Gastrointest. Endosc. **89**(6), AB560 (2019).
11. N. Muguruma, H. Miyamoto, T. Okahisa, and T. Takayama, "Endoscopic molecular imaging: status and future perspective," Clin. Endosc. **46**(6), 603–610 (2013).
12. W. Du, N. Rao, D. Liu, H. Jiang, C. Luo, Z. Li, and B. Zeng, "Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images," Ieee Access. **7**, 142053 (2019).
13. S. Menon and N. Trudgill, "How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis," Endosc. Int. Open **02**(02), E46–E50 (2014).
14. Y. Gao, Z. D. Zhang, S. Li, Y. T. Guo, Q. Y. Wu, S. H. Liu, and Y. Lu, "Deep neural network-assisted computed tomography diagnosis of metastatic lymph nodes from gastric cancer," Chin. Med. J. **132**(23), 2804–2811 (2019).
15. S. Wang, Y. Zhu, L. Yu, H. Chen, H. Lin, X. Wan, and P. A. Heng, "RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification," Med. Image Anal. **58**, 101549 (2019).
16. H. Ali, M. Yasmin, M. Sharif, and M. H. Rehmani, "Computer assisted gastric abnormalities detection using hybrid texture descriptors for chromoendoscopy images," Compu. Methods Programs Biomed. **157**, 39–47 (2018).
17. T. Kanesaka, T. C. Lee, N. Uedo, K. P. Lin, H. Z. Chen, and J. Y. Lee, . . . & H. T. Chang, "Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging," Gastrointestinal endoscopy. **87**(5), 1339–1344 (2018).
18. Y. Mori, S. ei Kudo, H. E. N. Mohmed, M. Misawa, N. Ogata, H. Itoh, M. Oda, and K. Mori, "Artificial intelligence and upper gastrointestinal endoscopy: current status and future perspective," Dig.Endosc. **31**(4), 378–388 (2019).
19. Y. Zhu, Q. C. Wang, M. D. Xu, Z. Zhang, J. Cheng, Y. S. Zhong, Y. Q. Zhang, W. F. Chen, L. Q. Yao, P. H. Zhou, and Q. L. Li, "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy," Gastrointest.Endosc. **89**(4), 806–815.e1 (2019).
20. F. Van Der Sommen, S. Zinger, E. J. Schoon, and P. H. De With, "Supportive automatic annotation of early esophageal cancer using local gabor and color features," Neurocomputing **144**, 92–106 (2014).
21. R. Zhou, C. Yang, M. Q. H. Meng, G. Xu, C. Hu, and B Li, "Capsule endoscopy images classification by random forests and ferns," in *4th IEEE International Conference on Information Science and Technology* (2014). pp. 414–417.
22. K. Shankar, Y. Zhang, Y. Liu, L. Wu, and C. H. Chen, "Hyperparameter tuning deep learning for diabetic retinopathy fundus image classification," IEEE Access **8**, 118164–118173 (2020).
23. H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, S. Mirjalili, and M. K. Khan, "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms," Proc. SPIE **11734**, 99–110 (2021).
24. B. Sun, Z. Wu, Y. Hu, and T. Li, "Golden subject is everyone: A subject transfer neural network for motor imagery-based brain computer interfaces," Neural Netw. **151**, 111–120 (2022).

25. B. Sun, Z. Liu, Z. Wu, C. Mu, and T. Li, “Graph convolution neural network based end-to-end channel selection and classification for motor imagery brain-computer interfaces,” IEEE Transactions on Industrial Informatics **19**(9), 9314–9324 (2023).
26. Z. Kong, T. Li, J. Luo, and S. Xu, “Automatic tissue image segmentation based on image processing and deep learning,” J. Healthcare Eng. **2019**, 1–10 (2019).
27. W. Xie, C. Jacobs, J. P. Charbonnier, and B. Van Ginneken, “Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans,” IEEE Trans. Med. Imaging **39**(8), 2664–2675 (2020).
28. J. Zilly, J. M. Buhmann, and D. Mahapatra, “Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation,” Computerized Medical Imaging and Graphics. **55**, 28–41 (2017).
29. R. Huang, W. Xie, and J. A. Noble, “VP-Nets: Efficient automatic localization of key brain structures in 3D fetal neurosonography,” Med. Image Anal. **47**, 127–139 (2018).
30. W. Sun, T. L. B. Tseng, J. Zhang, and W. Qian, “Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data,” Comput. Med. Imaging Graph. **57**, 4–9 (2017).
31. A. J. de Groof, M. R. Struyvenberg, J. van der Putten, F. van der Sommen, K. N. Fockens, W. L. Curvers, and J. J. Bergman, “Deep-learning system detects neoplasia in patients with Barrett’s esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking,” Gastroenterol. **158**(4), 915–929.e4 (2020).
32. M. Ohmori, R. Ishihara, K. Aoyama, K. Nakagawa, H. Iwagami, N. Matsuura, and T. Tada, “Endoscopic detection and differentiation of esophageal lesions using a deep neural network,” Gastrointest. Endosc. **91**(2), 301–309.e1 (2020).
33. T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, and T. Tada, “Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images,” Gastric Cancer **21**(4), 653–660 (2018).
34. Y. Zhang, F. Li, F. Yuan, K. Zhang, L. Huo, Z. Dong, and L. Shen, “Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence,” Dig. Liver Dis. **52**(5), 566–572 (2020).
35. L. Li, Y. Chen, Z. Shen, X. Zhang, J. Sang, Y. Ding, and C. Yu, “Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging,” Gastric Cancer. **23**(1), 126–132 (2020).
36. X. Zhang, W. Hu, F. Chen, J. Liu, Y. Yang, L. Wang, H. Duan, and J. Si, “Gastric pre- cancerous diseases classification using CNN with a concise model,” PLoS One **12**(9), e0185508 (2017).
37. C. Wang, Y. Li, J. Yao, B. Chen, J. Song, and X Yang, “Localizing and identifying intestinal metaplasia based on deep learning in oesophagoscope,” in *8th International Symposium on Next Generation Electronics (ISNE)* (2019), pp. 1–4.
38. T. Yan, P. K. Wong, I. C. Choi, C. M. Vong, and H. H. Yu, “Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images,” Computers in Biology and Medicine. **126**, 104026 (2020).
39. Y. Horiuchi, K. Aoyama, Y. Tokai, T. Hirasawa, S. Yoshimizu, A. Ishiyama, and T. Tada, “Convolutional neural network for differentiating gastric cancer from gastritis using magnified endoscopy with narrow band imaging,” Dig. Dis. Sci. **65**(5), 1355–1363 (2020).
40. X. Liu, C. Wang, J. Bai, and G. Liao, “Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images,” Neurocomputing **392**, 253–267 (2020).
41. B. J. Cho, C. S. Bang, S. W. Park, Y. J. Yang, S. I. Seo, H. Lim, and G. H. Baik, “Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network,” Endosc. **51**(12), 1121–1129 (2019).
42. T. K. Lui, K. K. Wong, L. L. Mak, E. W. To, V. W. Tsui, Z. Deng, and W. K. Leung, “Feedback from artificial intelligence improved the learning of junior endoscopists on histology prediction of gastric lesions,” Endosc. Int. Open **08**(02), E139–E146 (2020).
43. J. Fu, H. Zheng, and T Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2017), pp. 4438–4446.
44. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, and H. Zhang, . . . & X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2017), pp. 3156–3164.
45. Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y Yang, “Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification,” arXiv, arXiv:1801.09927 (2018).
46. W. Du, N. Rao, C. Dong, Y. Wang, D. Hu, L. Zhu, and T. Gan, “Automatic classification of esophageal disease in gastroscopic images using an efficient channel attention deep dense convolutional neural network,” Biomed. Opt. Express **12**(6), 3066–3081 (2021).
47. U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, “A survey of decision fusion and feature fusion strategies for pattern classification,” IETE Tech. Rev. **27**(4), 293–307 (2010).
48. Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K Barnard, “Attentional feature fusion,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), pp. 3560–3569.
49. F Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2017), pp. 1251–1258.
50. C. Szegedy, S. Ioffe, V. Vanhoucke, and A Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proceedings of the AAAI conference on artificial intelligence* (2017), (Vol. 31, No. 1).

Biomedical Optics EXPRESS

51. F. Yu and V Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv, arXiv:1511.07122 (2015).
52. Y. Zhang, B. Kang, B. Hooi, S. Yan, and J Feng, "Deep long-tailed learning: A survey," arXiv, arXiv:2110.04596 (2021).
53. T. Li, Y. Shang, and W. Ge, "Optical technologies for healthcare and wellness applications," J. Healthcare Eng. **2019**, 1321348 (2019).
54. W. Drexler and J. G. Fujimoto, "State-of-the-art retinal optical coherence tomography," Progress Retinal Eye Res. **27**(1), 45–88 (2008).
55. S. Bhat, I. V. Larina, K. V. Larin, M. E. Dickinson, and M. Liebling, "4D reconstruction of the beating embryonic heart from two orthogonal sets of parallel optical coherence tomography slice-sequences," IEEE Trans. Med. Imaging **32**(3), 578–588 (2013).
56. N. D. Gladkova, G. A. Petrova, N. K. Nikulin, S. G. Radenska-Lopovok, L. B. Snopova, Y. P. Chumakov, and F. I. Feldchtein, "In vivo optical coherence tomography imaging of human skin: norm and pathology," Skin Res. Technol. **6**(1), 6–16 (2000).
57. L. Wang, Y. Yang, J. Li, W. Tian, K. He, T. Xu, and T. Li, "Automatic classification of gastric lesions in gastroscopic images using a lightweight deep learning model with attention mechanism and cost-sensitive learning," Front. Phys. **10**, 1033422 (2022).